Data preparation

Dataset: 2020 NFL Predictions

2020 NFL predictions data preparation:

This project focuses on the preparation and cleaning of a dataset related to NFL predictions for the year 2020 using R markdown. The dataset is sourced from Elo ratings and includes various statistics related to NFL teams and games.

Step 1: There are two sheets in this dataset, Let's ensure that the columns are of the same datatype before combining

Mismatched columns:

```
##                      Column Sheet1_Type Sheet2_Type
## playoff             playoff     logical   character
## elo1_pre           elo1_pre   character     numeric
## elo2_post         elo2_post   character     numeric
## qb1_value_pre qb1_value_pre     numeric   character
## score1               score1     numeric   character
## score2               score2     numeric   character
```

Step 2: There are a few mismatched columns as listed above, let's correct them

Data types of the mismatched columns after data type conversion:

```
##          Column Sheet1_Type Sheet2_Type
## 1      elo1_pre     numeric     numeric
## 2     elo2_post     numeric     numeric
## 3 qb1_value_pre     numeric     numeric
## 4        score1     numeric     numeric
## 5        score2     numeric     numeric
```

Step 3: Date formatting

The dates are in numeric format, let's convert them to date format. Date column before conversion:

```
## [1] 44084 44087 44087 44087 44087 44087
```

Date column after conversion:

```
##
## 1905-07-12 2020-09-10 2020-09-13 2020-09-14 2020-09-17 2020-09-20 2020-09-21
##          1          1         13          2          1         14          1
## 2020-09-24 2020-09-27 2020-09-28 2020-10-01 2020-10-04 2020-10-05 2020-10-08
##          1         14          1          1         12          2          1
## 2020-10-11 2020-10-12 2020-10-13 2020-10-18 2020-10-19 2020-10-22 2020-10-25
##         11          1          1         12          2          1         11
## 2020-10-26 2020-10-29 2020-11-01 2020-11-02 2020-11-05 2020-11-08 2020-11-09
##          1          1         12          1          1         12          1
## 2020-11-12 2020-11-15 2020-11-16 2020-11-19 2020-11-22 2020-11-23 2020-11-26
##          1         12          1          1         12          1          2
## 2020-11-29 2020-11-30 2020-12-02 2020-12-06 2020-12-07 2020-12-08 2020-12-10
##         12          1          1         12          2          1          1
## 2020-12-13 2020-12-14 2020-12-17 2020-12-19 2020-12-20 2020-12-21 2020-12-25
##         14          1          1          2         12          1          1
## 2020-12-26 2020-12-27 2020-12-28 2021-01-03 2021-01-09 2021-01-10 2021-01-16
##          3         11          1         16          3          3          2
## 2021-01-17 2021-01-24 2021-02-07
##          2          2          1
```

Step 4: Outlier detection

Removing a row with an unusual date (1905-07-12) as it is irrelevant in a 2020 NFL prediction dataset

```
##
## 2020-09-10 2020-09-13 2020-09-14 2020-09-17 2020-09-20 2020-09-21 2020-09-24
##          1         13          2          1         14          1          1
## 2020-09-27 2020-09-28 2020-10-01 2020-10-04 2020-10-05 2020-10-08 2020-10-11
##         14          1          1         12          2          1         11
## 2020-10-12 2020-10-13 2020-10-18 2020-10-19 2020-10-22 2020-10-25 2020-10-26
##          1          1         12          2          1         11          1
## 2020-10-29 2020-11-01 2020-11-02 2020-11-05 2020-11-08 2020-11-09 2020-11-12
##          1         12          1          1         12          1          1
## 2020-11-15 2020-11-16 2020-11-19 2020-11-22 2020-11-23 2020-11-26 2020-11-29
##         12          1          1         12          1          2         12
## 2020-11-30 2020-12-02 2020-12-06 2020-12-07 2020-12-08 2020-12-10 2020-12-13
##          1          1         12          2          1          1         14
## 2020-12-14 2020-12-17 2020-12-19 2020-12-20 2020-12-21 2020-12-25 2020-12-26
##          1          1          2         12          1          1          3
## 2020-12-27 2020-12-28 2021-01-03 2021-01-09 2021-01-10 2021-01-16 2021-01-17
##         11          1         16          3          3          2          2
## 2021-01-24 2021-02-07
##          2          1
```

Step 5: Season correction

All values of ‘season’ should be 2020 since this prediction was done on a 2020 dataset.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       2    2020    2020    1998    2020    2020
```

Not all values are 2020. Lets correct them.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2020    2020    2020    2020    2020    2020
```

Step 6: Handing missing values in 'playoff' column and removing the 'neutral' column

BEFORE

```
## [1] "Frequency of playoff column:"
```

```
##
##    c    d    s    w <NA>
##    2    4    1    6  255
```

```
## [1] "Frequency of neutral column:"
```

```
##
##   0   1   9  99
## 264   1   1   2
```

playoff: labeling the missing values as so it is easy to analyze results
neutral: let's flag and remove the 'neutral' column because except for 3 errored values (9,99,1) and others are 0, meaning they are home/away games.
AFTER:

```
## [1] "Frequency of playoff column\n"
```

```
##
##    c    d    s    w <NA>
##    2    4    1    6  255
```

Step 7: Handling missing values, identifying and rectifying outliers in the rest of the columns

Column: team 1

```
##
##  ARI  ATL  BAL  BUF  CAR  CHI  CIN  CLE  DAL  DEN  DET   GB  HOU  IND  JAX   KC
##    7    8    8   10    8    8    8    8    8    8    9    7    8    8   10
##  LAC  LAR  MIA  MIN   NE   NO  NYG  NYJ  OAK  PHI  PIT  SEA   SF   TB  TEN  WSH
##    8    7    8    8    8   10    8    8    8    8    9    9    8    8    9    9
## <NA>
##    4
```

```
## # A tibble: 4 × 2
##   team1 qb1
##   <chr> <chr>
## 1 <NA>  Kyler Murray
## 2 <NA>  Aaron Rodgers
## 3 <NA>  John Wolford
## 4 <NA>  <NA>
```

There are 4 rows with NA. Let's remove the row where all values are empty. Let's find and impute the team names from the quaterbacks respective to these rows.

```
##
## ARI ATL BAL BUF CAR CHI CIN CLE DAL DEN DET  GB HOU IND JAX  KC LAC LAR MIA MIN
##   8   8   8  10   8   8   8   8   8   8   8  10   7   8   8  10   8   8   8   8
##  NE  NO NYG NYJ OAK PHI PIT SEA  SF  TB TEN WSH
##   8  10   8   8   8   8   9   9   8   8   9   9
```

Column: team 2

```
##
##     ARI     ATL     BAL     BUF     CAR     CHI     CIN     CLE     DAL     DEN
##       8       8      10       9       8       8       8      10       8       8
##     DET      GB     HOU Houston     IND     JAX      KC     LAC     LAR     MIA
##       8       7       7       1       9       8       8       8      10       7
##     MIN      NE      NO     NYG     NYJ     OAK OAKLAND     PHI     PIT     SEA
##       8       8       8       8       8       7       1       7       8       8
##      SF      TB     TEN     WSH    <NA>
##       8      11       8       8       3
```

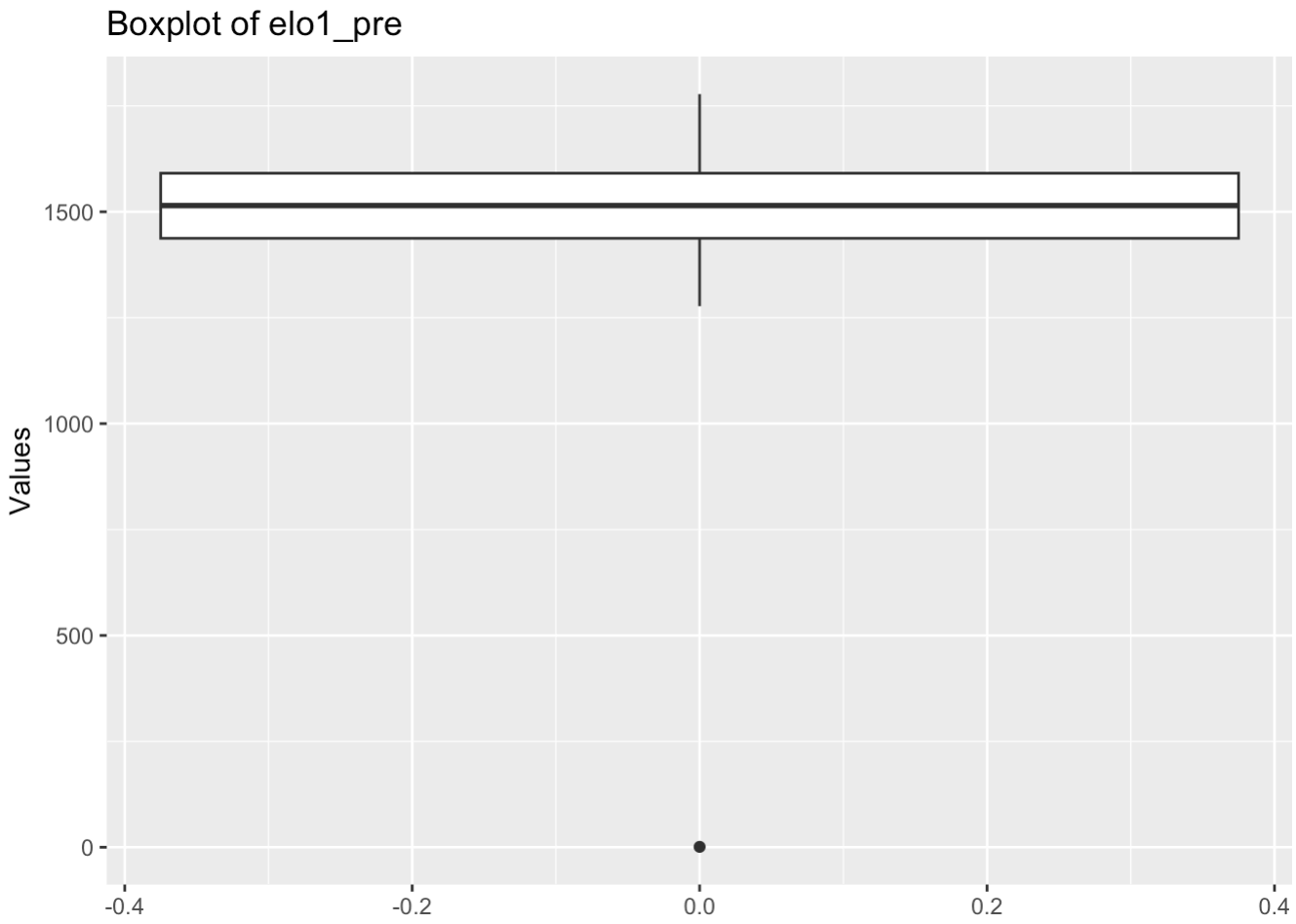
```
## # A tibble: 3 × 2
##   team2 qb2
##   <chr> <chr>
## 1 <NA>  Mitchell Trubisky
## 2 <NA>  Carson Wentz
## 3 <NA>  Tua Tagovailoa
```

1 - Three missing values, let's deal with them as we did for 'team 1' column 2 - Abbreviations of Houston, and oakland has to be corrected

```
##
## ARI ATL BAL BUF CAR CHI CIN CLE DAL DEN DET  GB HOU IND JAX  KC LAC LAR MIA MIN
##   8   8  10   9   8   9   8  10   8   8   8   7   8   9   8   8   8  10   8   8
##  NE  NO NYG NYJ OAK PHI PIT SEA  SF  TB TEN WSH
##   8   8   8   8   8   8   8   8   8  11   8   8
```

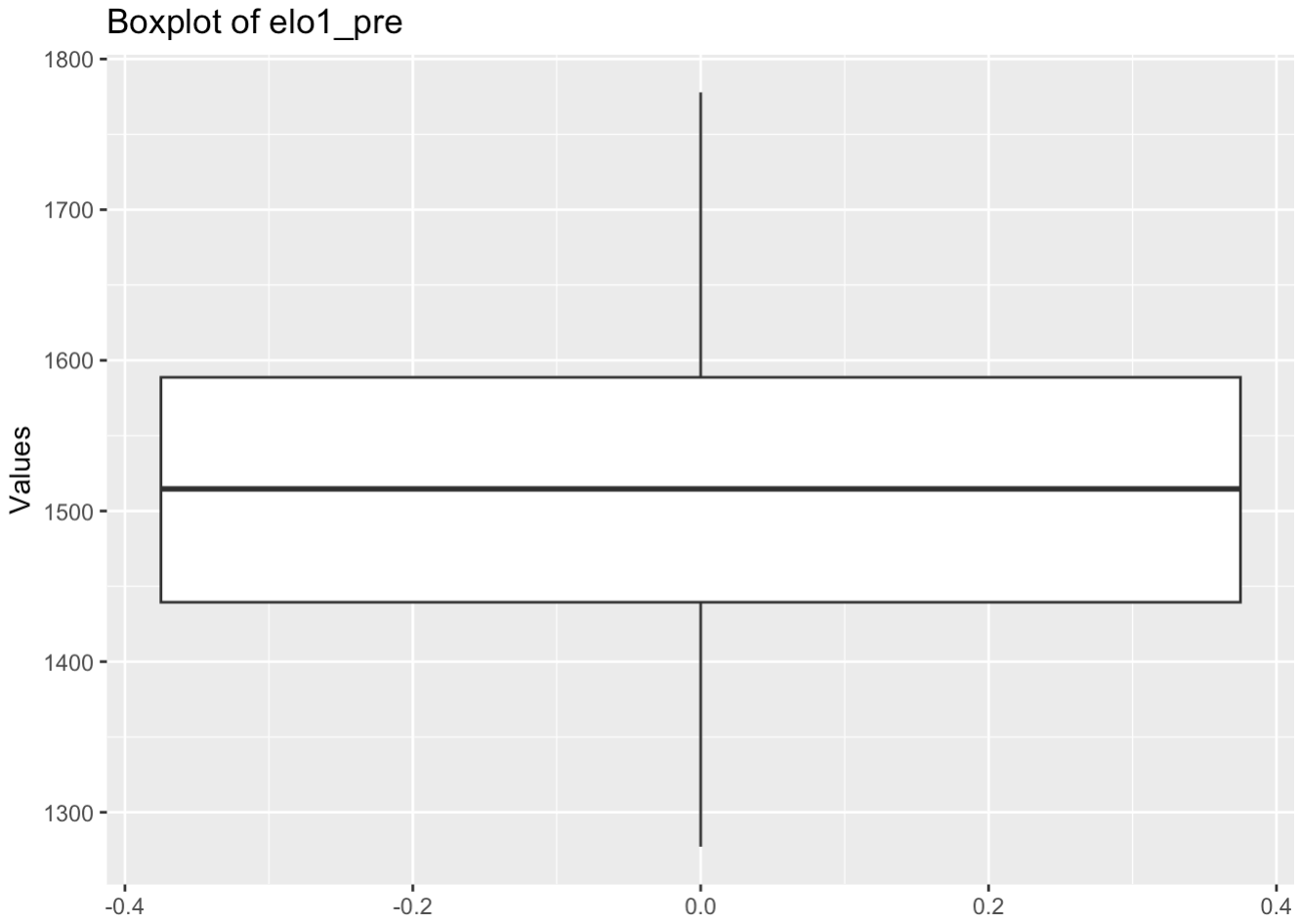
Column: elo1_pre

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       1    1437    1515    1510    1591    1778       4
```



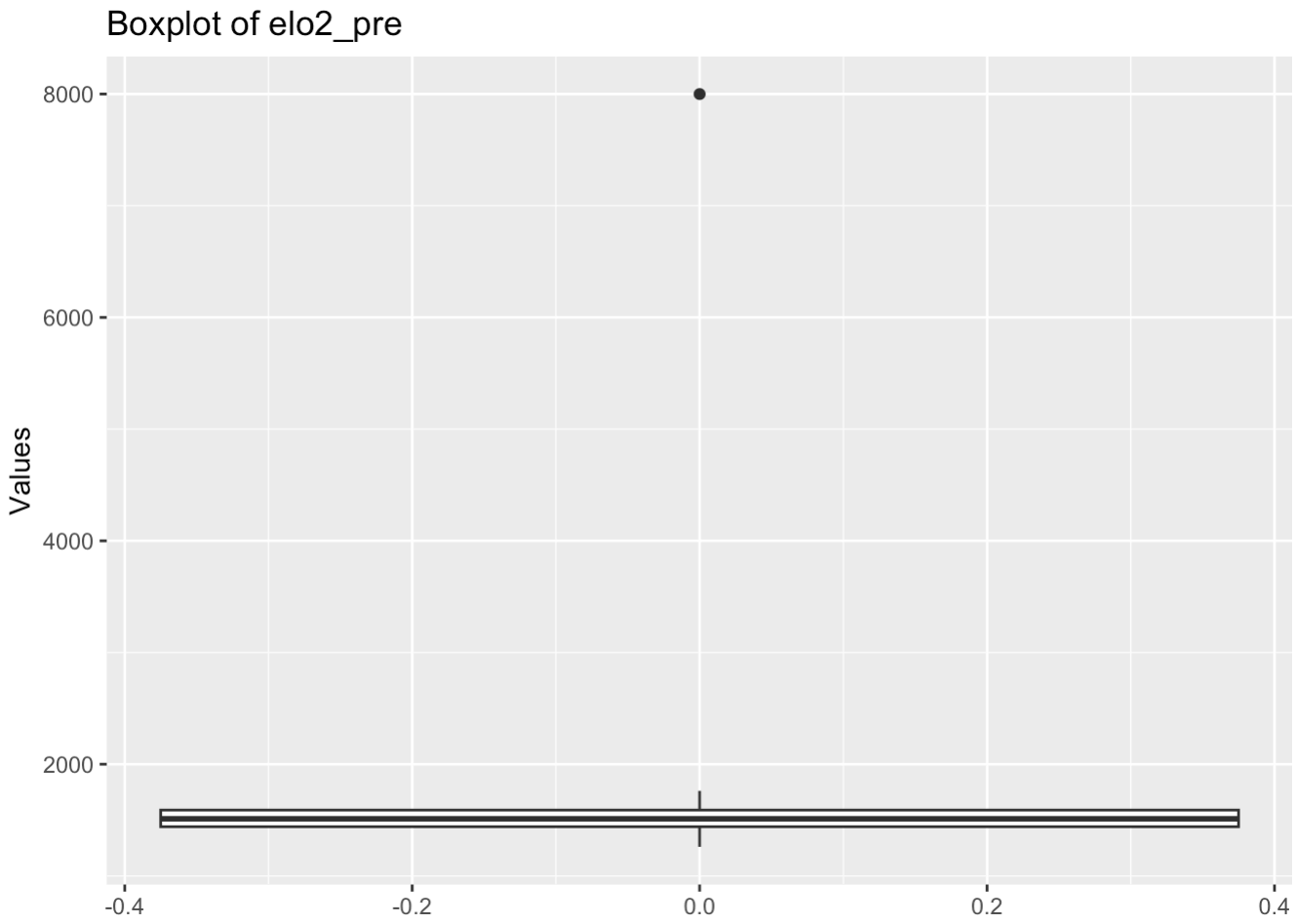
50% of the data falls within a range close to 1500 while the rest ranging from around 1200 to 1800. There is one outlier around 0, represented by the dot. Let's impute the missing values and the outlier with the median value. Since we have outliers, it is best to impute with median instead of mean

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1277    1439    1515    1515    1589    1778
```



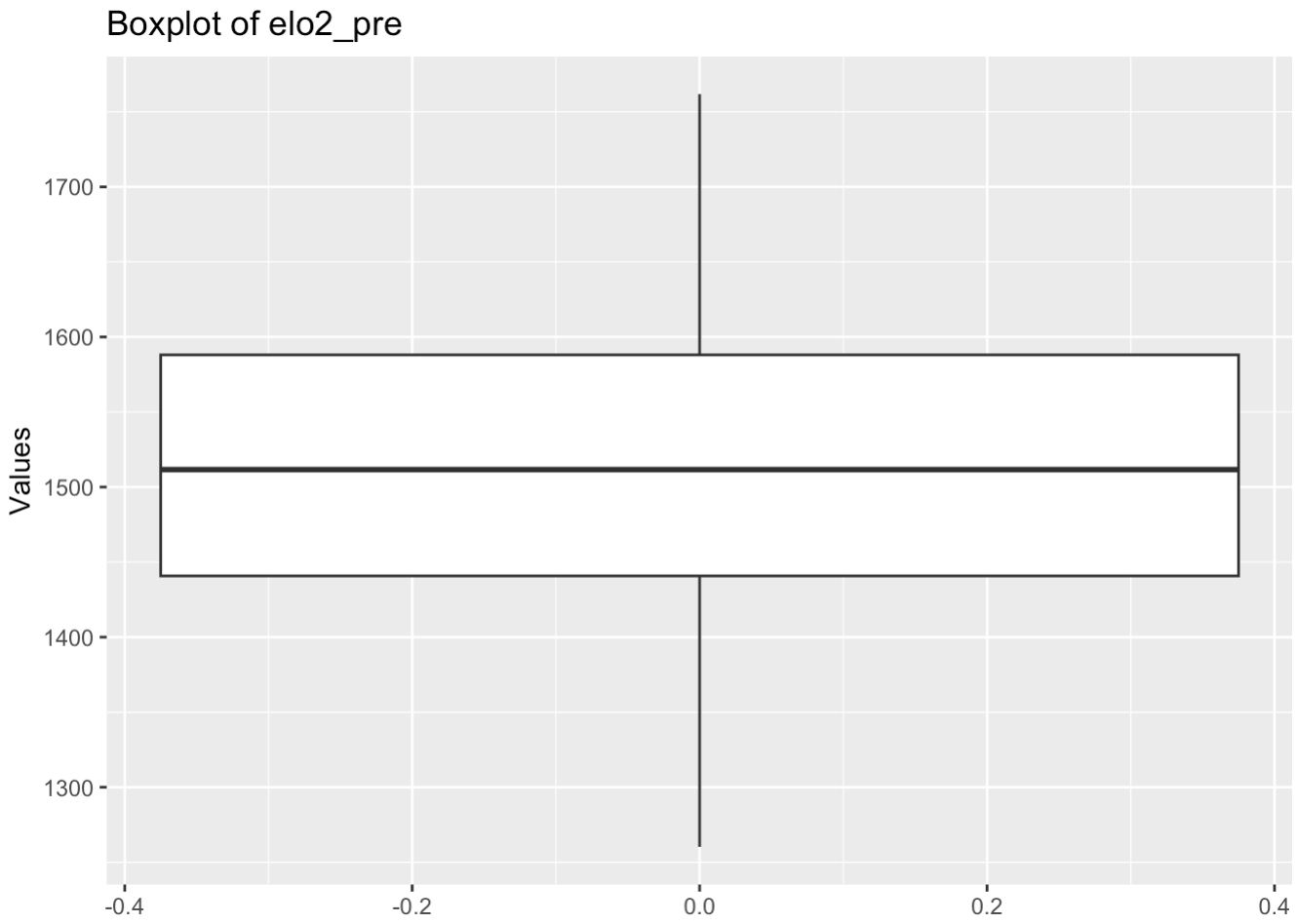
Column: elo2_pre

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1260    1441    1512    1533    1589    8000       2
```



Its the opposite with this column, there is an outlier around 8000 and 2 missing values.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1260    1441    1512    1509    1588    1762
```



Columns: elo_prob1, and elo_prob2

```
##
## Summary for elo_prob1 :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.1720  0.4665  0.5993  0.5869  0.7071  0.9370       2
##
## Summary for elo_prob2 :
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    0.063    0.293    0.401   15.506    0.534 4000.000        2
```

Upper bound outlier found in elo_prob2 (Max - 4000). Replace values of elo_prob2 with (1-elo_prob1) according to the formula

```
##
## Summary for elo_prob1 :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1720  0.4669  0.5993  0.5870  0.7067  0.9370
##
## Summary for elo_prob2 :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06298 0.29325 0.40065 0.41302 0.53313 0.82805
```

Columns: elo1_post, and elo2_post

```
##
## Summary for elo1_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       1    1430    1514    1497    1586    1778       4
##
## Summary for elo2_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   -1504    1435    1510    1500    1591    1775       7
```

elo1_post - outlier at 0, it should be the value 1 (min) and two missing values. elo2_post - negative outlier, elo ratings are never negative. Let's impute it with median

```
##
## Summary for elo1_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1260    1438    1514    1509    1585    1778
##
## Summary for elo2_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1256    1437    1510    1512    1590    1775
```

Columns: qbelo1_pre, and qbelo2_pre

```
##
## Summary for qbelo1_pre :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0    1438    1514    1497    1580    1757       4
##
## Summary for qbelo2_pre :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1259    1441    1516    1512    1581    1742       5
```

qbelo1_pre - Quarterback Elo ratings, a value of 0 is highly unlikely and unrealistic. It looks like an error, so, lets impute them with median qbelo2_pre - Missing values

```
##
## Summary for qbelo1_pre :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1272    1446    1514    1514    1579    1757
##
## Summary for qbelo2_pre :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1259    1442    1516    1512    1579    1742
```

Columns: qb1, and qb2

```
## # A tibble: 4 × 2
##   qb1   team1
##   <chr> <chr>
## 1 <NA>  TB
## 2 <NA>  CAR
## 3 <NA>  JAX
## 4 <NA>  KC
```

```
## # A tibble: 1 × 2
##   qb2   team2
##   <chr> <chr>
## 1 <NA>  IND
```

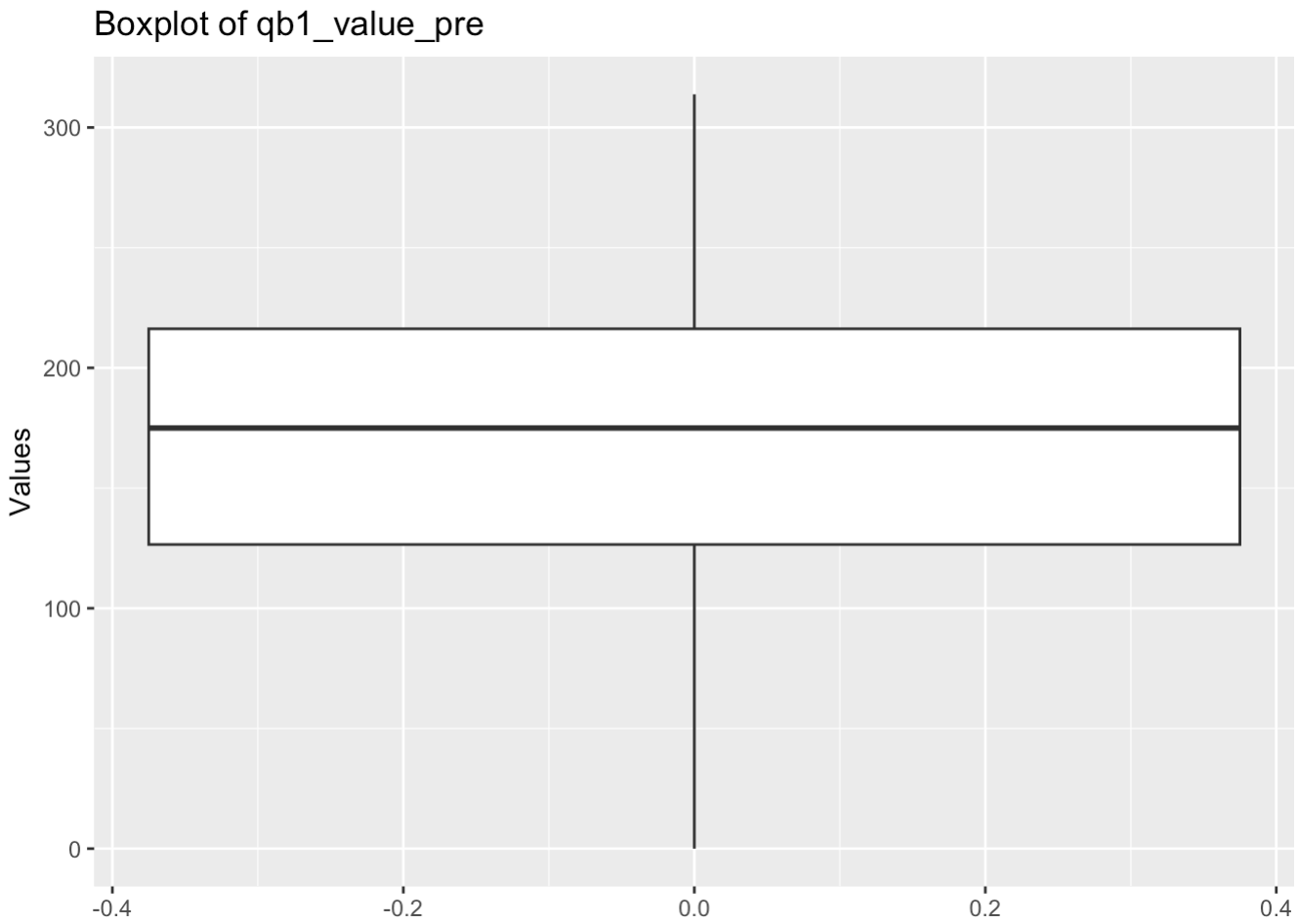
Let's find and update the quaterback names from the teams we found and remove the row where all values are NA. Also, scan for spelling mistakes or special characters. None found

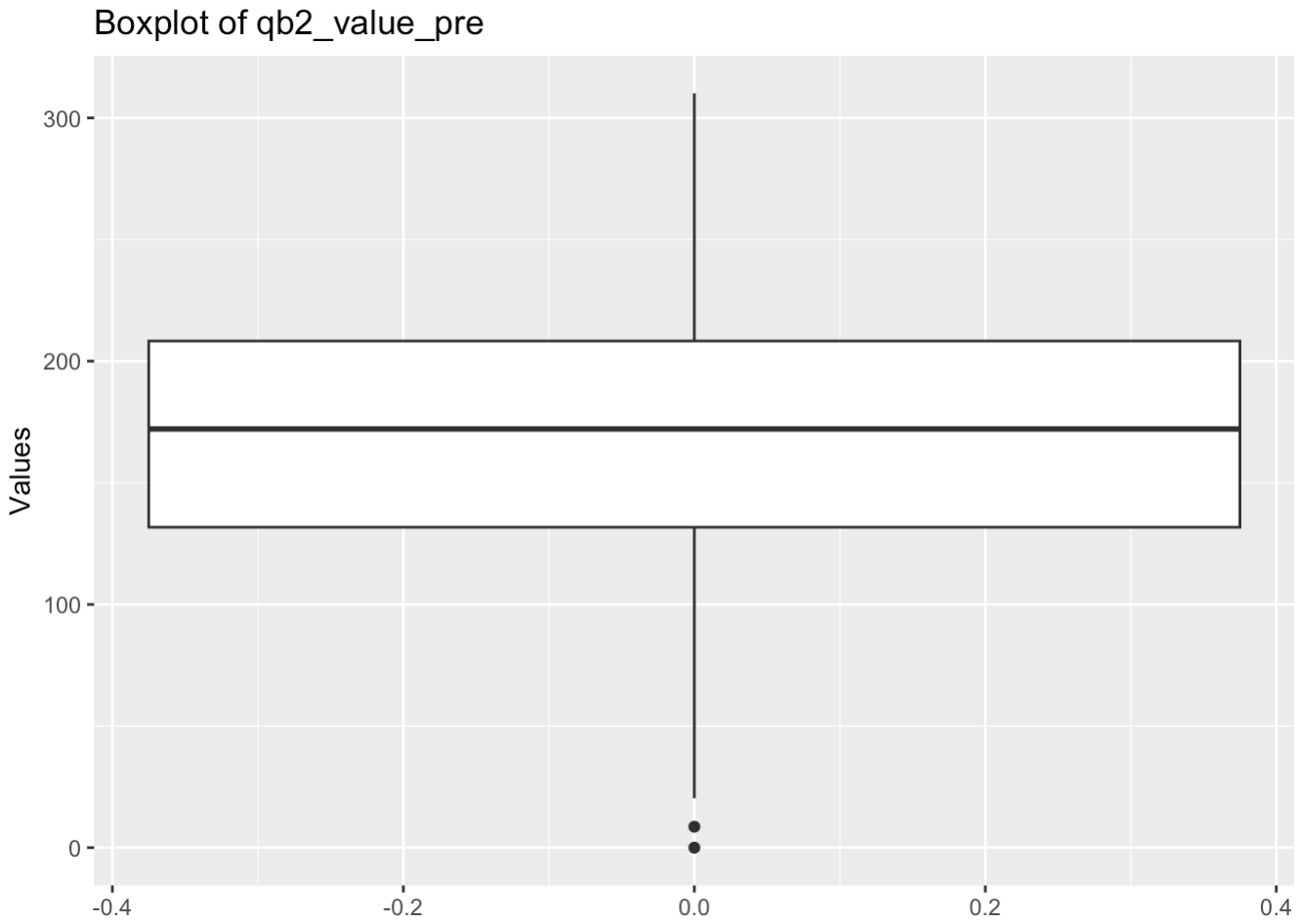
```
##
##      Aaron Rodgers         Alex Smith        Andy Dalton     Baker Mayfield
##                 10                  1                  4                  8
## Ben Roethlisberger      Brandon Allen      C.J. Beathard         Cam Newton
##                  9                  3                  1                  8
##       Carson Wentz         Chad Henne         Colt McCoy       Dak Prescott
##                  6                  1                  1                  3
##       Daniel Jones         Derek Carr     Deshaun Watson         Drew Brees
##                  7                  8                  7                  9
##          Drew Lock     Dwayne Haskins    Gardner Minshew    Garrett Gilbert
##                  6                  4                  4                  1
##         Jake Luton        Jalen Hurts         Jared Goff       Jeff Driskel
##                  1                  2                  7                  1
##    Jimmy Garoppolo         Joe Burrow         Joe Flacco       John Wolford
##                  3                  4                  2                  1
##         Josh Allen     Justin Herbert     Kendall Hinton       Kirk Cousins
##                 10                  8                  1                  8
##         Kyle Allen       Kyler Murray      Lamar Jackson          Matt Ryan
##                  3                  8                  8                  8
##   Matthew Stafford       Mike Glennon  Mitchell Trubisky         Nick Foles
##                  8                  3                  4                  4
##       Nick Mullens        P.J. Walker    Patrick Mahomes      Philip Rivers
##                  4                  1                  9                  8
##     Russell Wilson        Ryan Finley   Ryan Fitzpatrick     Ryan Tannehill
##                  9                  1                  3                  9
##        Sam Darnold    Taylor Heinicke        Taysom Hill  Teddy Bridgewater
##                  6                  1                  1                  7
##          Tom Brady     Tua Tagovailoa
##                  8                  5
```

```
##
##      Aaron Rodgers         Alex Smith        Andy Dalton     Baker Mayfield
##                  7                  5                  5                 10
##        Ben DiNucci Ben Roethlisberger      Brandon Allen       Brett Rypien
##                  1                  7                  2                  1
##        Brian Hoyer      C.J. Beathard         Cam Newton       Carson Wentz
##                  1                  1                  7                  6
##         Colt McCoy       Dak Prescott       Daniel Jones         Derek Carr
##                  1                  2                  7                  8
##     Deshaun Watson         Drew Brees          Drew Lock     Dwayne Haskins
##                  8                  5                  7                  2
##    Gardner Minshew         Jake Luton        Jalen Hurts         Jared Goff
##                  5                  1                  2                  9
##    Jimmy Garoppolo         Joe Burrow         Joe Flacco       John Wolford
##                  3                  6                  2                  1
##         Josh Allen           Justin H     Justin Herbert       Kirk Cousins
##                  9                  1                  6                  8
##         Kyle Allen       Kyler Murray      Lamar Jackson      Mason Rudolph
##                  1                  8                  9                  1
##          Matt Ryan   Matthew Stafford       Mike Glennon  Mitchell Trubisky
##                  8                  8                  2                  6
##         Nick Foles       Nick Mullens    Patrick Mahomes      Philip Rivers
##                  3                  4                  8                  9
## Robert Griffin III     Russell Wilson   Ryan Fitzpatrick     Ryan Tannehill
##                  1                  8                  4                  8
##        Sam Darnold        Taysom Hill  Teddy Bridgewater          Tom Brady
##                  6                  3                  8                 11
##     Tua Tagovailoa       Tyrod Taylor
##                  4                  1
```

Columns: qb1_value_pre, and qb2_value_pre

```
##
## Summary for qb1_value_pre :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   126.5   175.0   170.6   216.2   313.8       4
##
## Summary for qb2_value_pre :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   131.7   172.1   169.8   208.3   310.1       5
```





potential outliers around 0, a realistic minimum value for a quarterback's Elo rating in established leagues should be above 100 or 150, as Elo ratings typically start around 1500.

```
##
## Summary for qb1_value_pre :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   102.2   150.2   175.0   184.5   216.1   313.8
##
## Summary for qb2_value_pre :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   102.0   145.7   172.1   180.2   207.8   310.1
```

Columns: qb1_adj, and qb2_adj

```
##
## Summary for qb1_adj :
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -242.488   -6.634    6.350   -1.731   15.692   54.827        8
##
## Summary for qb2_adj :
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -218.569   -4.025    6.120   -1.464   16.223   53.096        3
```

Both the adjusted values have outliers and NAs. let's find the lower and upper bounds by calculating IQR. And, replace the outliers < lower bound | outliers > upper bound | NAs with median

```
##
## Summary for qb1_adj :
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -30.4603  -0.8247   6.3501   7.6494  15.3847  48.9537
##
## Summary for qb2_adj :
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -34.3734  -0.3922   6.1197   8.0197  15.7942  45.8382
```

Columns qbelo_prob1, and qbelo_prob2

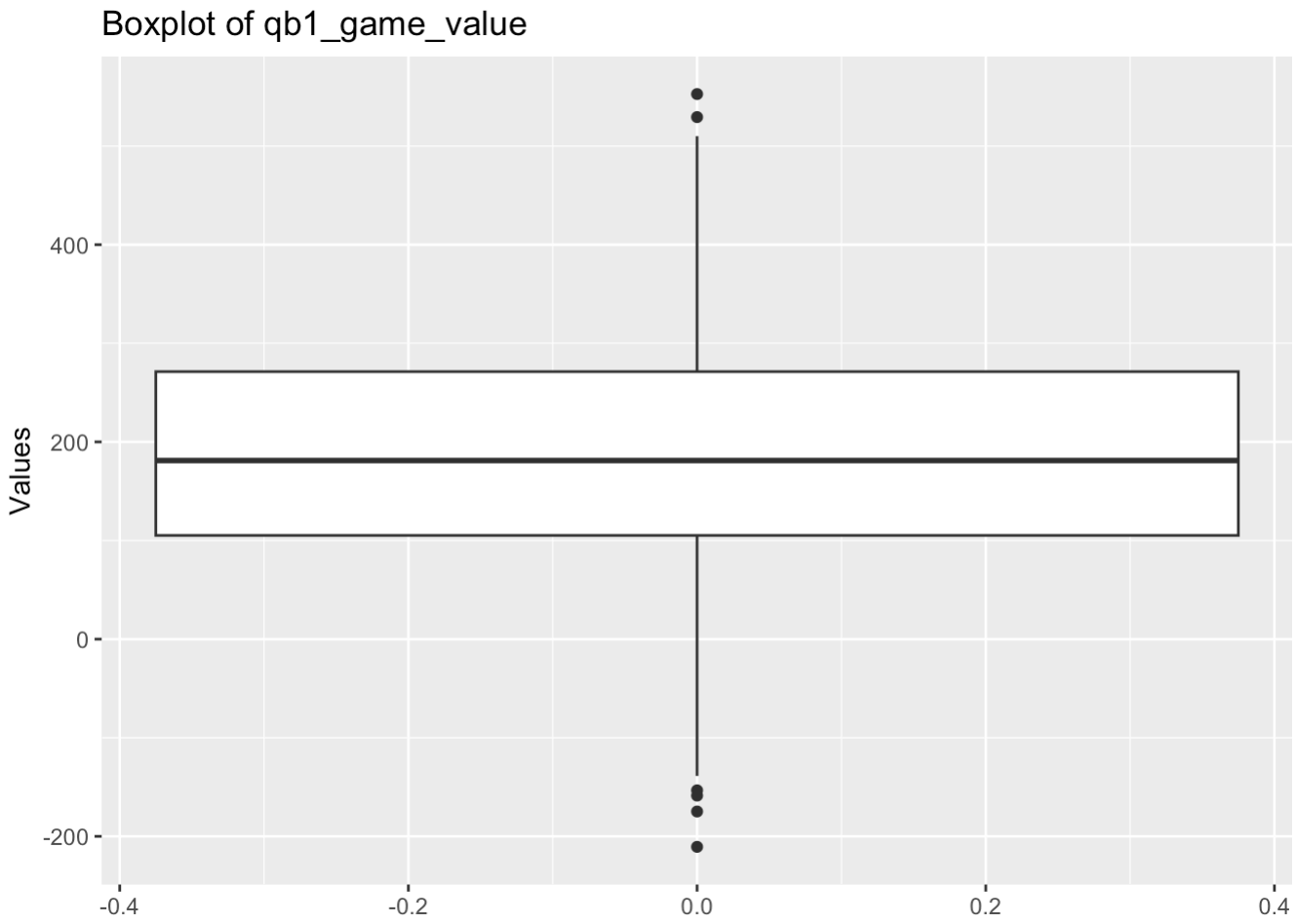
```
##
## Summary for qbelo_prob1 :
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##  0.07023  0.40905  0.54902  0.64160  0.69645 25.00000        3
##
## Summary for qbelo_prob2 :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.06693 0.30426 0.45370 0.45142 0.59095 0.92977       3
```

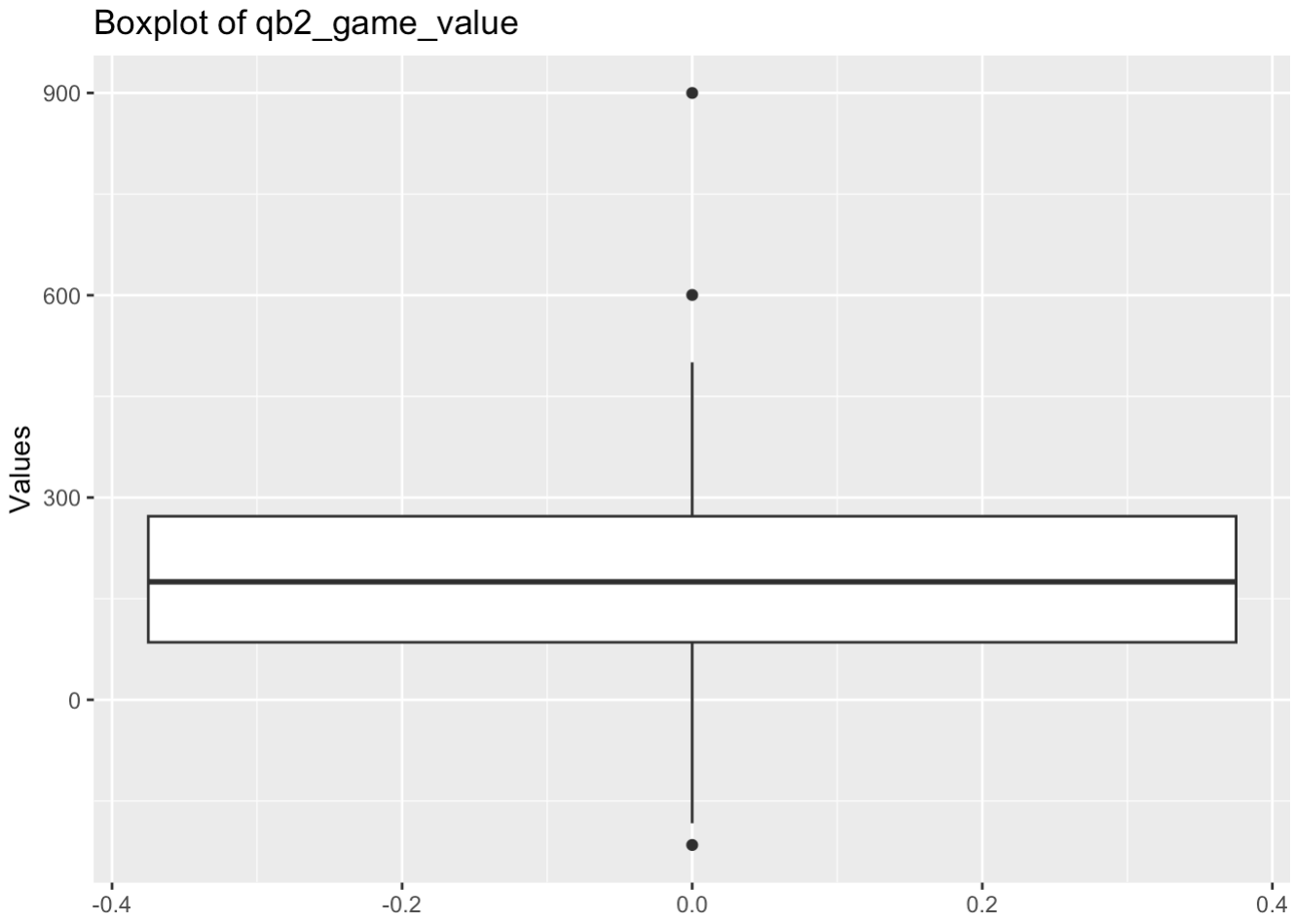
Both have 3 NAs each, qbelo_prob1 alone has an outlier (25)

```
##
## Summary for qbelo_prob1 :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07023 0.40969 0.54902 0.54898 0.69531 0.93307
##
## Summary for qbelo_prob2 :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06693 0.30469 0.45370 0.45144 0.59031 0.92977
```

Columns: qb1_game_value, and qb2_game_value

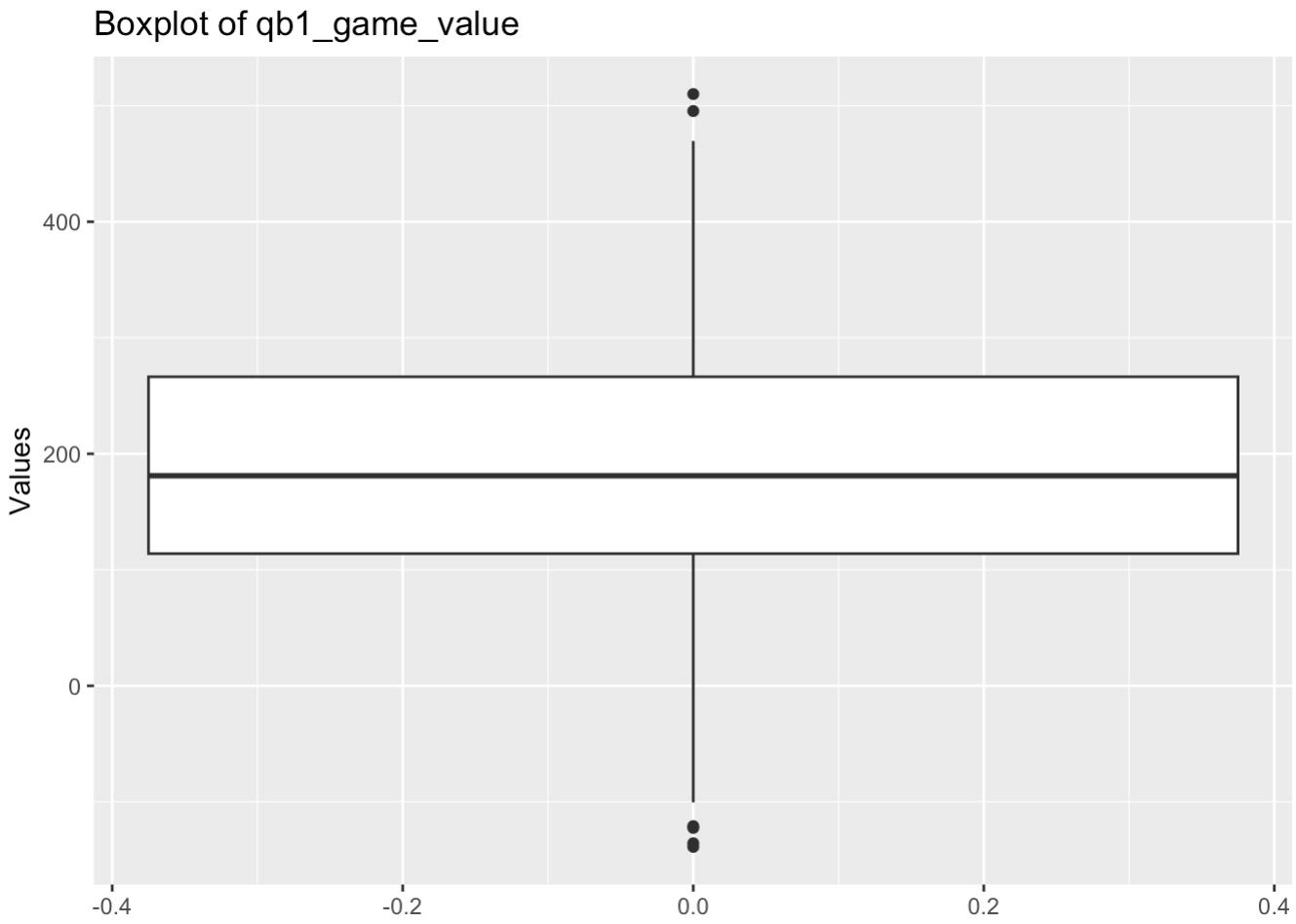
```
##
## Summary for qb1_game_value :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -210.7   105.2   181.1   181.5   271.3   552.8       5
##
## Summary for qb2_game_value :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -215.31   85.33  174.96  179.26  272.24  900.00       5
```

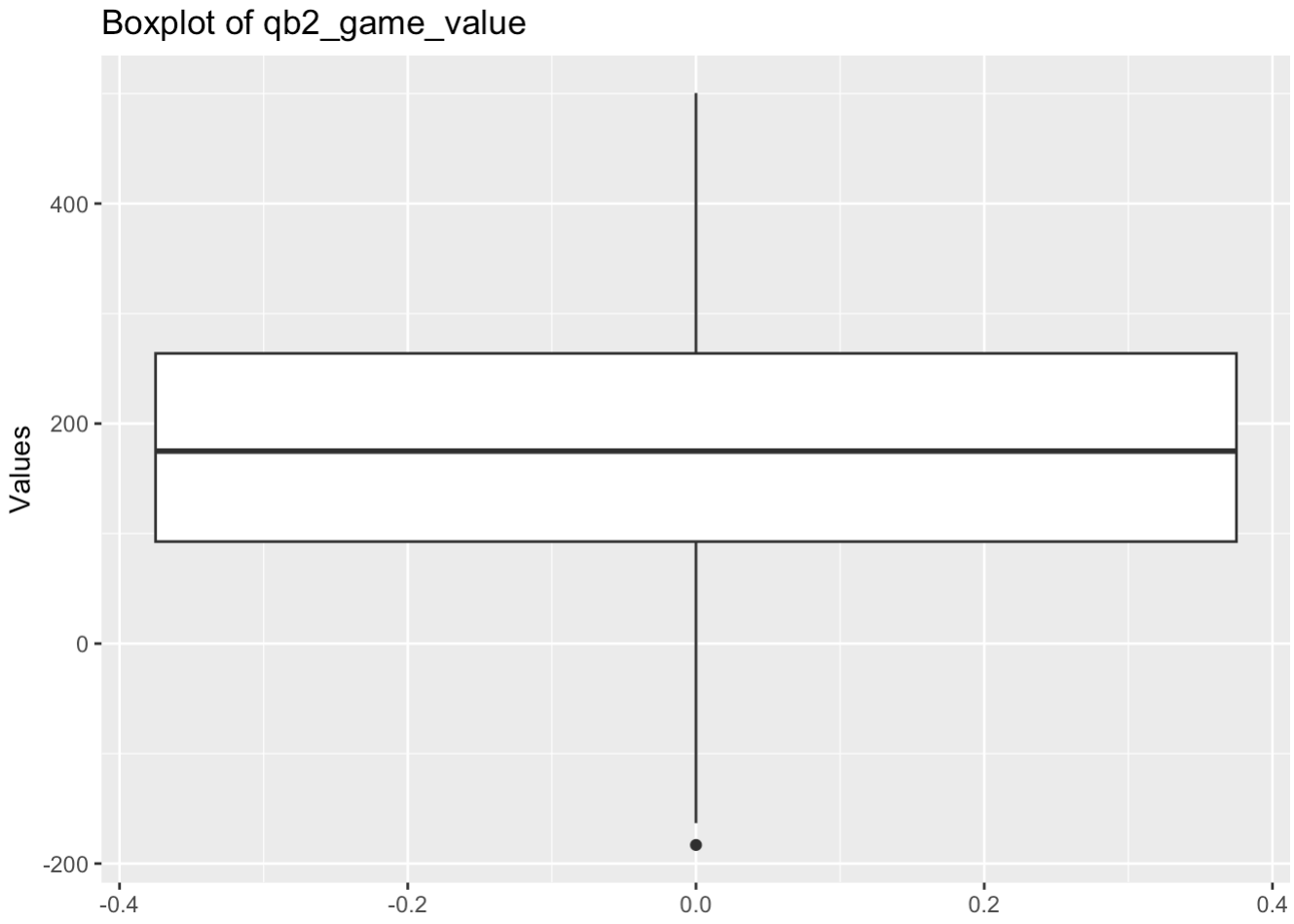




Both have positive and negative outliers. Let's call the calculate_bounds function to find the upper and lower bounds, then replace them with median

```
##
## Summary for qb1_game_value :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -138.8   113.9   181.1   184.1   266.4   510.1
##
## Summary for qb2_game_value :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -183.10   92.73  174.96  176.33  263.80  500.55
```





In prediction models, especially for sports like NFL, outliers may represent extreme scenarios, such as unusually bad or good performances predicted for a quarterback. These could be rare but realistic outcomes. If the outliers are natural and don't heavily impact your analysis, it's perfectly acceptable to leave them in the dataset.

Columns: qb1_game_value, qb2_game_value, qbelo1_post, and qbelo2_post

```
##
## Summary for qb1_value_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -13.57  131.24  175.25  171.51  216.35  310.13       5
##
## Summary for qb2_value_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   5.574 128.308 173.248 169.674 211.612 313.828       5
##
## Summary for qbelo1_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      15    1437    1515    1507    1584    1757       5
##
## Summary for qbelo2_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1255    1444    1511    1513    1582    1755       5
```

qb1_value_post - negative outlier qbelo1_post - outlier around 0

```
##
## Summary for qb1_value_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.431 133.957 175.249 175.125 215.067 310.131
##
## Summary for qb2_value_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.574 128.778 173.248 169.741 210.006 313.828
##
## Summary for qbelo1_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1259    1441    1515    1513    1583    1757
##
## Summary for qbelo2_post :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1255    1444    1511    1513    1580    1755
```

Columns: score 1 and score 2

```
##
## Summary for score1 :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -16.00   19.00   24.00   24.67   31.00   99.00       4
##
## Summary for score2 :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   17.00   25.00   24.62   31.00   49.00       4
```

Score 1 - Negative scores are considered outliers as they are unrealistic, clearly errors. Also, a score of 99 is extremely unlikely

```
##
## Summary for score1 :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   19.50   24.00   24.65   31.00   56.00
##
## Summary for score2 :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   17.00   25.00   24.63   31.00   49.00
```

```
## The cleaned dataset has been saved as: cleaned_dataset.csv
```